ABSTRACT OF THE DISCLOSURE

A method for automatically detecting table data in a document that is described by a page definition language and converting the table data into a markup language representation. The document may have one or more pages. The page definition language description of the document provides a list of words, the position of the each on a page with respect to a predetermined reference point, and the size of each word. The present invention automatically identifies table data in the document by utilizing one or more table-identifying features. A first table-identifying feature may be the number of word clusters on a line. A second table-identifying feature may be the vertical alignment of word clusters between lines. A third table-identifying feature may be the changes in text density or space density between lines.